Margin Maximization and Implicit Bias

Ziniu Li
ziniuli@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, Shenzhen, China

April 15, 2021

Mainly based on Chapter 15 of the deep learning theory lecture notes by Matus Telgarsky.

Outline

Background

Separability and Margin Maximization

Gradient Flow Maximizes Margins of Linear Predictors

Smoothed Margins Are Nondecreasing For Homogeneous Functions

Motivation

- Deep neural networks perform well, even though parameters norms are large and there is no explicit regularization [Neyshabur et al., 2015, Zhang et al., 2017].
- To understand this issue, there have been efforts toward the implicit bias by gradient descend and related optimization algorithms [Soudry et al., 2018, Ji and Telgarsky, 2019, Ji et al., 2020, Gunasekar et al., 2018, Arora et al., 2019].
- In particular, [Soudry et al., 2018] showed that gradient descend on the cross-entropy (and the exp) loss is implicitly biased towards a maximum margin direction for linearly separable data.

Motivation

- Margin maximization of first-order methods applied to exponentially-tailed losses was first proved for coordinate descend [Telgarsky, 2013] for Adaboost.
  - An introductory material for this topic is `https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec10.pdf`.
- The main idea is that the empirical risk after the monotone transformation $\ln(\cdot)$ is $\ln\sum\exp(\cdot)$, which is similar to $\max(\cdot)$; hence it is closely related to margin maximization.

Notation

- Assume that we are provided $(x_i, y_i)_{i=1}^n$, the unnormalized margin mapping is defined by
$$m_i(w) = y_i f(x_i; w). \tag{1}$$
- By this choice, the unnormalized risk $\mathcal{L}$ is defined as
$$\mathcal{L}(w) = \sum_i \ell\left(m_i(w)\right) = \sum_i \ell\left(y_i f(x_i; w)\right). \tag{2}$$
We will use the exponential loss $\ell(z) = \exp(-z)$.
- We assume that f is locally Lipschitz and L-homogeneous in w.
  - f is locally Lipschitz when for every point x, there exists a neighborhood S such that $\{x\} \subset S$ and f is Lipschitz when restricted to S.
  - f is L-homogeneous if $f(cx) = c^L f(x)$.

Outline

Background

Separability and Margin Maximization

Gradient Flow Maximizes Margins of Linear Predictors

Smoothed Margins Are Nondecreasing For Homogeneous Functions

Separability and Margin Maximization: Linear Predictor

- Consider a linear predictor $x \mapsto \langle w, x\rangle$, by "separable", we mean that y_i agrees with the direction sgn$(\langle w, x_i\rangle)$.
- Let us introduce the concept of strict separability:
$$\min_i y_i\langle w, x_i\rangle > 0.$$
- It seems reasonable, or a nice inductive bias, if we are far from 0 possible:
$$\max_{w\in ?} \min_i y_i\langle w, x_i\rangle > 0.$$
Here "?" indicates that we must somehow normalize w; otherwise, the above quantity could go to $+\infty$.

Separability and Margin Maximization: Linear Predictor

Definition 1 (Linearly separable, maximum margin).

Data is linearly separable when there exists $w \in \mathbb{R}^d$ so that $\min_i y_i \langle w, x_i \rangle > 0$. In this situation, the (ℓ_2) maximum margin predictor (which is unique!) is given by

$$\bar{u} := \operatorname*{argmax}_{\|w\|=1} \min_i y_i \langle w, x_i \rangle. \tag{3}$$

And the margin is

$$\gamma := \min_i y_i \langle \bar{u}, x_i \rangle. \tag{4}$$

Separability and Margin Maximization

- We want to consider more general cases beyond linear predictors.
- An easy extension is that $f(x; w)$ is L-homogeneous.
- But, we need to check that whether definitions like margin are still well-behaved.

Separability and Margin Maximization: L-homogeneous

Proposition 1.

Suppose $f(x;w)$ is L-homogeneous in w, ℓ is the exponential loss, and there exists $\widehat{w}$ with

$$\widehat{\mathcal{R}}(\widehat{\omega}) < \frac{\ell(0)}{n} = \frac{1}{n}. \tag{5}$$

Then $\inf_w \widehat{\mathcal{R}}(w) = 0$, and the infimum is not attained. Here $\widehat{\mathcal{R}}$ means the empirical risk, i.e.,

$$\widehat{\mathcal{R}}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(y_i f(x_i;w)\right).$$

Proof of Proposition 1

We first show that (5) implies the margin is large than 0.

$$\max_i \ell\left(m_i(\widehat{w})\right) \leq \sum_{i=1}^{n} \ell\left(m_i(\widehat{w})\right) = n\widehat{\mathcal{R}}(\widehat{w}) < \ell(0),$$

thus applying ℓ^{-1} to both sides yields $\min_i m_i(\widehat{w}) > 0$. Therefore,

$$\begin{aligned} 0 \leq \inf_w \widehat{\mathcal{R}}(w) \leq \limsup_{c\to\infty} \widehat{\mathcal{R}}(c\widehat{w}) &\leq \sum_{i=1}^{n} \limsup_{c\to\infty} \ell\left(m_i(c\widehat{w})\right) \\ &= \sum_{i=1}^{n} \limsup_{c\to\infty} \ell\left(c^L m_i(\widehat{w})\right) = 0. \end{aligned}$$

Hence, we get $\inf_w \widehat{\mathcal{R}}(w) = 0$.

Separability and Margin Maximization: L-homogeneous

- It seems wired that how can we “find” an “optimum” when solutions at infinity by Proposition 1?
- By L-homogeneous of f, we have

$$\min_i m_i(w) := \|w\|^L \min_i m_i\left(\frac{w}{\|w\|}\right).$$

- Therefore we could build a better-behaved “margin” by normalizing the originally defined margin by $\|w\|^L$:

$$\gamma(w) := \min_i m_i\left(\frac{w}{\|w\|}\right) = \min_i \frac{m_i(w)}{\|w\|^L}.$$

Separability and Margin Maximization: L-homogeneous

- Let us introduce the smoothed margin [Lyu and Li, 2020, Schapire and Freund, 2012]:

$$\tilde{\gamma}(w) := \frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L}. \tag{6}$$

Recall that $\ell^{-1}(\mathcal{L}(w))$ has what we have mentioned "log-sum-exp" structure.

- To understand this, we note that

$$\frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L} \le \frac{\ell^{-1}\left(\max_i \ell(m_i(w))\right)}{\|w\|^L} = \frac{\min_i m_i(w)}{\|w\|^L},$$
$$\frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L} + \frac{\ln n}{\|w\|^L} = \frac{\ell^{-1}\left(\sum_i \ell(m_i(w)/n)\right)}{\|w\|^L} \ge \frac{\min_i m_i(w)}{\|w\|^L}.$$

That is,

$$\frac{\min_i m_i(w)}{\|w\|^L} - \frac{\ln n}{\|w\|^L} \le \frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L} \le \frac{\min_i m_i(w)}{\|w\|^L}. \tag{7}$$

Separability and Margin Maximization: L-homogeneous

Definition 2 (margin, maximum margin, smoothed margin).

Say the data is $\vec{m}$-separable when there exists w so that $\min_i m_i(w) > 0$. Define the margin, maximum margin, and smoothed margin respectively as

$$\gamma(w) := \min_i m_i\left(\frac{w}{\|w\|}\right) = \min_i \frac{m_i(w)}{\|w\|^L},$$

$$\bar{\gamma} := \max_{\|w\|=1} \gamma(w),$$

$$\widetilde{\gamma}(w) := \frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L}.$$

Separability and Margin Maximization: L-homogeneous

Proposition 2.

Suppose data is $\vec{m}$-separable (i.e., there exists w so that $\min_i m_i(w) > 0$). Then

- $\overline{\gamma} := \max_{\|w\|=1} \gamma(w)$ *is well-defined (i.e., the maximum is attained).*
- *For any $w \neq 0$, we have,*

  $$\lim_{c\to\infty} \widetilde{\gamma}(cw) = \gamma(w).$$

  In particular, for any $\widehat{w}$ satisfying $\overline{\gamma} = \gamma(\widehat{w})$, $\lim_{c\to\infty} \widetilde{\gamma}(c\widehat{w}) = \overline{\gamma}$.

Proof of Proposition 2

- The first part follows by continuity of $m_i(w)$ and compactness of $\{w \in \mathbb{R}^d : \|w\| = 1\}$.
- The second part uses (7):

$$\begin{aligned}
& \frac{\min_i m_i(w)}{\|w\|^L} - \frac{\ln n}{\|w\|^L} \leq \frac{\ell^{-1}(\mathcal{L}(w))}{\|w\|^L} \leq \frac{\min_i m_i(w)}{\|w\|^L} \\
\implies \quad & \frac{\min_i m_i(cw)}{\|cw\|^L} - \frac{\ln n}{\|cw\|^L} \leq \frac{\ell^{-1}(\mathcal{L}(cw))}{\|cw\|^L} \leq \frac{\min_i m_i(cw)}{\|cw\|^L} \\
\implies \quad & \lim_{c\to\infty} \frac{\min_i m_i(w)}{\|w\|^L} - \frac{\ln n}{\|cw\|^L} \leq \lim_{c\to\infty} \frac{\ell^{-1}(\mathcal{L}(cw))}{\|cw\|^L} \leq \lim_{c\to\infty} \frac{\min_i m_i(w)}{\|w\|^L} \\
\implies \quad & \gamma(w) = \frac{\min_i m_i(w)}{\|w\|^L} = \lim_{c\to\infty} \frac{\ell^{-1}(\mathcal{L}(cw))}{\|cw\|^L} = \lim_{c\to\infty} \tilde{\gamma}(cw).
\end{aligned}$$

Outline

Background

Separability and Margin Maximization

Gradient Flow Maximizes Margins of Linear Predictors

Smoothed Margins Are Nondecreasing For Homogeneous Functions

Gradient Flow Maximizes Margins of Linear Predictors

- In this part, we consider linear predictors (i.e., 1-homogeneous).
- Recall the max-margin predictor and maximum margin defined in (3) and (4), respectively:

$$\bar{u} := \operatorname*{argmax}_{\|w\|=1} \min_i y_i \langle w, x_i \rangle, \quad \gamma := \min_i y_i \langle \bar{u}, x_i \rangle.$$

Lemma 1.

Consider a linear predictor: $x \mapsto \langle w, x \rangle$*, with linearly separable data and the exponential loss, and* $\max_i \|x_i y_i\| \le 1$*. In addition, assume* $w(0) = 0$*, consider the gradient flow:*

$$\dot{w}(t) = -\nabla \mathcal{L}(w(t)) \tag{8}$$

Then,

$$\mathcal{L}(w_t) \le \frac{1 + \ln(2nt\gamma^2)}{2t\gamma^2}, \tag{9}$$

$$\|w_t\| \ge \ln(2tn\gamma^2) - \ln\left(1 + \ln(2tn\gamma^2)\right). \tag{10}$$

Proof of Lemma 1

Theorem 1 (Theorem 10.4 of [Telgarsky, 2021]).

For any $z \in \mathbb{R}^d$, if $\widehat{\mathcal{R}}$ is convex and smooth, gradient flow satisfies

$$\widehat{\mathcal{R}}(w(t)) - \widehat{\mathcal{R}}(z) \le \frac{1}{2t}\left(\|w(0) - z\|^2 - \|w(t) - z\|^2\right).$$

Let $z = \ln(c)\bar{u}/\gamma$ for some $c > 0$, we have

$$\mathcal{L}\left(w(t)\right) \le \mathcal{L}(z) + \frac{1}{2t}\left(\|z\|^2 - \|w(t) - z\|^2\right) \le \sum_i \ell(m_i(z)) + \frac{\|z\|^2}{2t}$$

$$\le \sum_i \exp\left(-\ln(c)\right) + \frac{\ln^2(c)}{2t\gamma^2} = \frac{n}{c} + \frac{\ln^2(c)}{2t\gamma^2}.$$

Choosing $c := 2tn\gamma^2$, we get (9) (it seems $\ln^2$ is missing in (9) ?).

Proof of Lemma 1

Now, we try to prove the lower bound of $\|w_t\|$ in (10). By our assumption of $\max_i \|x_i y_i\| \leq 1$, we have

$$\begin{aligned} & |m_i(w_t)| = |y_i \langle w_t, x_i \rangle| \leq \|y_i x_i\| \|w\| \leq \|w\|, \\ \implies \quad & \max_i m_i(w_t) \leq \|w\|. \end{aligned}$$

Thus,

$$\ell\left(\|w_t\|\right) \leq \min_i \ell\left(m_i(w_t)\right) \leq \frac{1}{n}\mathcal{L}(w_t) \leq \frac{1+\ln^2(2tn\gamma^2)}{2tn\gamma^2}.$$

Applying $\ell^{-1}(z) = -\ln(z)$ on both sides, we obtain (10).

Gradient Flow Maximizes Margins of Linear Predictors

Theorem 2 (Margin maximization of linear predictors).

Consider a linear predictor: $x \mapsto \langle w, x \rangle$*, with linearly separable data and the exponential loss, and* $\max_i \|x_i y_i\| \leq 1$*. Then*

$$\gamma(w_t) \geq \widetilde{\gamma}(w_t) \geq \bar{\gamma} - \frac{\ln n}{\ln t + \ln(2n\gamma^2) - 2 \ln \ln(2tne\gamma^2)}. \tag{11}$$

♠ : The first inequality directly follows (7).

Proof of Theorem 2

For convenience, define

$$u(t) := \ell^{-1}\left(\mathcal{L}(w(t))\right), \quad v(t) := \|w(t)\|.$$

In this way, we get that

$$\widetilde{\gamma}\left(w(t)\right) := \frac{\ell^{-1}\left(\mathcal{L}(w)\right)}{\|w\|} = \frac{u(t)}{v(t)} = \frac{u(0)}{v(t)} + \frac{\int_0^t \dot{u}(s)ds}{v(t)}. \tag{12}$$

Our goal is to lower bound the second term.

Proof of Theorem 2

Note that $\ell' = -\ell$, thus

$$\dot{u}(t) = \left\langle \frac{-\nabla \mathcal{L}(w(t))}{\mathcal{L}(w(t))}, \dot{w}(t) \right\rangle \overset{(8)}{=} \frac{\|\dot{w}(t)\|^2}{\mathcal{L}((w(t))}, \tag{13}$$

$$v(t) = \|w(t) - w(0)\| = \left\| \int_0^t \dot{w}(s) ds \right\| \leq \int_0^t \|\dot{w}(s)\| \, ds \tag{14}$$

In addition,

$$\begin{aligned} \|\dot{w}(s)\| \geq \langle \dot{w}(s), \bar{u} \rangle &= \left\langle -\sum_i x_i y_i \ell'(m_i(w(s))), \bar{u} \right\rangle \\ &= \sum_i \ell\left(m_i(w(s))\right) \langle x_i y_i, \bar{u} \rangle \geq \gamma \sum_i \ell\left(m_i(w(s)\right) = \gamma \mathcal{L}(w) \end{aligned} \tag{15}$$

Proof of Theorem 2

Combining previous inequalities, we have

$$\begin{aligned}\frac{1}{v(t)}\int_0^t \dot{u}(s)ds &\overset{(13)}{\geq} \frac{1}{v(t)}\int_0^t \frac{\|\dot{w}(s)\|^2}{\mathcal{L}(w(s))}ds \overset{(14)}{\geq} \frac{1}{\int_0^t \|\dot{w}(s)\|\,ds}\int_0^t \frac{\|\dot{w}(s)\|^2}{\mathcal{L}(w(s))}ds \\ &\overset{(15)}{\geq} \frac{\gamma}{\int_0^t \|\dot{w}(s)\|\,ds}\int_0^t \|\dot{w}(s)\|\,ds \\ &= \gamma.\end{aligned}$$

Back to (12), we obtain that

$$\widetilde{\gamma}(w(t)) = \frac{u(0)}{v(t)} + \frac{1}{v(t)}\int_0^t \dot{u}(s)ds \geq \frac{u(0)+\gamma}{v(t)}.$$

Finally, note that $u(0) = \ell^{-1}(\mathcal{L}(0)) = -\ln n$, we get the desired result:

$$\frac{u(0)}{v(t)} = \frac{-\ln(n)}{\|w(t)\|} \overset{(10)}{\geq} \frac{-\ln(n)}{\ln(t)+\ln(2n\gamma^2) - 2\ln\ln(2tne\gamma^2)}.$$

Outline

Background

Separability and Margin Maximization

Gradient Flow Maximizes Margins of Linear Predictors

Smoothed Margins Are Nondecreasing For Homogeneous Functions

Smoothed Margins Are Nondecreasing For Homogeneous Functions

- In the nonlinear case, we do not have a general result, but instead only prove that the smoothed margins are nondecreasing.

Theorem 3 (Originally from [Lyu and Li, 2020], simplification due to [Ji et al., 2020]). *Suppose there exists* t_0 *with* $\widetilde{\gamma}(w(t_0)) > 0$. *Then* $t \mapsto \widetilde{\gamma}(w(t))$ *is nondecreasing along* $[t_0, \infty)$.

Proof of Theorem 3

Recall our previous notations:

$$u_t := \ell^{-1}(\mathcal{L}(w_t)), \quad v_t := \|w_t\|^L .$$

So that we can write $\widetilde{\gamma}$ as

$$\widetilde{\gamma}_t := \widehat{\gamma}(w(t)) = \frac{u_t}{v_t}.$$

By the quotient rule,

$$\frac{d}{dt}\widetilde{\gamma}_t = \frac{\dot{u}_t v_t - \dot{v}_t u_t}{v_t^2}.$$

Therefore it suffices to show that $v_t \neq 0$ and that the numerator is nonnegative.

♠ : we need a lower bound on $\dot{u}_t$ and an upper bound on $\dot{v}_t$.

Proof of Theorem 3

We will use the following property of L-homogeneous functions.

Lemma 2 (Lemma 14.2 of [Telgarsky, 2021]).

Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ *is locally Lipschitz continuous and* L*-positive homogeneous. For any* $w \in \mathbb{R}^d$ *and* $s \in \partial f(w)$*,*

$$\langle s, w \rangle = Lf(w). \tag{16}$$

Note a technical fact that $\ell' = -\ell$,

$$\begin{aligned} \langle w, \dot{w} \rangle &= \sum_j -\ell'(m_j(w)) \langle w, \nabla m_j(w) \rangle \overset{(16)}{=} L \sum_j -\ell'(m_j(w)) m_j(w) \\ &= L \sum_j -\ell'(m_j(w)) \ell^{-1}\left(\ell(m_j(w))\right) \geq L \sum_j -\ell'(m_j(w)) \ell^{-1}\left(\mathcal{L}(w)\right) \\ &= L\mathcal{L}(w) \ell^{-1}\left(\mathcal{L}(w)\right). \end{aligned} \tag{17}$$

Proof of Theorem 3

Back to our goal of $\dot{v}_t$ and $\dot{v}_t$,

$$\dot{v}_t = \frac{d}{dt}\langle w_t, w_t\rangle^{L/2} = \frac{L}{2}\langle w_t, w_t\rangle^{L/2-1}2\langle w_t, \dot{w}_t\rangle = L\left\|w_t\right\|^{L-2}\langle w_t, \dot{w}_t\rangle. \tag{18}$$

Consequently,

$$\dot{v}_t = L\left\|w_t\right\|^{L-1}\left\langle \frac{w_t}{\|w_t\|}, \dot{w}_t\right\rangle \le L\left\|w_t\right\|^{L-1}\sup_{\|v\|\le 1}\langle v, \dot{w}\rangle = L\left\|w_t\right\|^{L-1}\left\|\dot{w}\right\|$$

For $\dot{u}_t$, using $\ell(z) = \exp(-z)$,

$$\dot{u}_t = -\frac{\langle L(w_t), \dot{w}_t\rangle}{\mathcal{L}(w_t)} \overset{(8)}{=} \frac{\left\|\dot{w}(t)\right\|^2}{\mathcal{L}((w(t))} \ge \frac{\left\|\dot{w}\right\|}{\mathcal{L}(w_t)\left\|w_t\right\|}\langle \dot{w}_t, w_t\rangle \overset{(17)}{\ge} \frac{L\left\|\dot{w}\right\|\ell^{-1}(\mathcal{L}(w))}{\left\|w_t\right\|}.$$

Combing the above inequalities,

$$\dot{u}_t v_t - \dot{v}_t u_t \ge \frac{L\left\|\dot{w}_t\right\|\ell^{-1}\left(\mathcal{L}(w)\right)}{\left\|w_t\right\|}\left\|w_t\right\|^L - L\left\|w_t\right\|^{L-1}\left\|\dot{w}_t\right\|\ell^{-1}\left(\mathcal{L}(w)\right) = 0.$$

Proof of Theorem 3

It remains to show that v_t is nonzero. First, note that $v_0 > 0$ since $\mathcal{L}(w_t) < \ell(0)/n \leq \mathcal{L}(0)$? As before,

$$\begin{aligned} \dot{v}_t &\overset{(18)}{=} L \left\| w_t \right\|^{L-2} \langle w_t, \dot{w}_t \rangle \overset{(17)}{\geq} L^2 \left\| w_t \right\|^{L-2} \mathcal{L}(w_t) \ell^{-1} \left(\mathcal{L}(w_t) \right) \\ &\overset{(6)}{=} L^2 \left\| w_t \right\|^{2L-2} \mathcal{L}(w_t) \widetilde{\gamma}_t . \end{aligned}$$

Let T be the first time where $v_t = 0$. For $t \in [0, T)$, $v_t > 0$ and thus $\widetilde{\gamma}_t \geq \widetilde{\gamma}_0$ and $\dot{v}_t > 0$, meaning such a time T cannot exist. Therefore, $\dot{v}_t > 0$ and $(d/dt)\widetilde{\gamma}_t > 0$.

References I

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In Advances in Neural Information Processing Systems 32, pages 7411–7422, 2019.

S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In Proceedings of the 35th International Conference on Machine Learning, pages 1827–1836, 2018.

Z. Ji and M. Telgarsky. Gradient descent aligns the layers of deep linear networks. In Proceedings of the 7th International Conference on Learning Representations, 2019.

Z. Ji, M. Dudík, R. E. Schapire, and M. Telgarsky. Gradient descent follows the regularization path for general losses. In Annual Conference on Learning Theory, pages 2109–2136, 2020.

K. Lyu and J. Li. Gradient descent maximizes the margin of homogeneous neural networks. In Proceedings of the 8th International Conference on Learning Representations, 2020.

References II

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In Workshop Track Proceedings of the 3rd International Conference on Learning Representations, 2015.

R. E. Schapire and Y. Freund. Boosting: Foundations and Algorithms. MIT Press, 2012.

D. Soudry, E. Hoffer, M. S. Nacson, S. Gunasekar, and N. Srebro. The implicit bias of gradient descent on separable data. Journal of Machine Learning Research, 19(70):1–57, 2018.

M. Telgarsky. Margins, shrinkage, and boosting. In Proceedings of the 30th International Conference on Machine Learning, pages 307–315, 2013.

M. Telgarsky. Deep learning theory lecture notes. https://mjt.cs.illinois.edu/dlt/, 2021. Version: 2021-02-14 v0.0-1dabbd4b (pre-alpha).

References III

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In Proceedings of the 5th International Conference on Learning Representations, 2017.